Neural-Network-Based Parameter Estimation in S-System Models of Biological Networks

Jonas S. Almeida **Eberhard O. Voit**
AlmeidaJ@MUSC.edu VoitEO@MUSC.edu

Department of Biometry and Epidemiology, Medical University of South Carolina,
Charleston, SC, 29425, U.S.A.

Abstract

The genomic and post-genomic eras have been blessing us with overwhelming amounts of data that are of increasing quality. The challenge is that most of these data alone are mere snapshots of the functioning organism and do not reveal the organizational structure of which the particular genes and metabolites are contributors. To gain an appreciation of their roles and functions within cells and organisms, genomic and metabolic data need to be integrated in systems models that allow the testing of hypotheses, generate experimentally testable predictions, and ultimately lead to true explanations. One type of data that is particularly well suited for such integration consists of time profiles, which show gene activities, metabolite concentrations, or protein prevalences at dense series of time points. We show with a specific example how such time series can be analyzed and evaluated, if some structural information about the data is available, even if this information is incomplete. The method consists of three components. The first is a particularly suitable mathematical modeling framework, namely *Biochemical Systems Theory*, in which parameters are direct indicators of the organization of the underlying phenomenon, the second is the training of an artificial neural network for data smoothing and complementation, and the third is a technique for reinterpreting differential equations in a fashion that facilitates parameter estimation. A prototype webtool for these analyses is available at `https://bioinformatics.musc.edu/webmetabol/`.

Keywords: artificial neural network, biochemical systems theory, genomics, metabolic profile, proteomics, S-system

1 Introduction

The hallmark of genomic and post-genomic research has been the high-throughput generation of data. Where it would constitute an entire doctoral dissertation to sequence a short piece of DNA a few decades ago, it is now hardly newsworthy if the sequencing of yet another genome is completed. Measuring the concentration of one or a few metabolites *in vivo* used to be a significant challenge, but it is now possible to measure hundreds or even thousands of small metabolites simultaneously with modern methods of mass spectrometry [3]. It used to be simply impossible to quantify immediate responses of cells to stimuli, but this can be done now with astonishing speed and accuracy [9, 16].

From the point of mathematical modeling, the abundant availability of data is dramatically changing the focus of research [20] and mandates the development of novel techniques. As a well-known example of such a shift, consider the analysis of microarray data, which requires statistical techniques that are diametrically opposite to established methods, because the number of (more or less) independent variables (expression levels of genes) is enormous, yet the sample sizes (repeats of expression measurements of the same gene) are typically small.

In this communication we address a different trend in data availability, which is bound to become more prevalent with novel developments in genomics, mass spectrometry, nuclear magnetic resonance, and proteomics techniques. This situation consists of time series of simultaneous measurements of many cellular components. In genomics, such time series are already available; an excellent example is the Stanford database [25], which contains time courses of gene expression in yeast after various stimuli. In the context of metabolic analysis, such "traces" could, for instance, be composed of concentration measurements of all glycolytic metabolites one, two, three, ..., twenty seconds or minutes after some stimulus; an example is [9]. In a proteomics setting, a trace could contain all (or many) protein prevalences over time, in response to changes in environmental milieu or throughout the manifestation of a disease.

The interesting challenge associated with these types of data is that they are direct reflections of the dynamics of a system at some level (genomic, gene regulatory, metabolic, proteomic, physiological), and if enough of this type of information is available, there is justified expectation that it provides clues about the true functioning of the system and may even reveal its full regulatory structure. It is well known that this information cannot be obtained solely from typical snapshot measurements in time. For instance, even a complete steady-state analysis does not provide information about the time scale of the system. Similarly, sensitivity analysis constrains the realm of possibilities for a given system, but it does not usually reveal the full regulatory structure (*e.g.*, [14]). The question then becomes how one may retrieve knowledge from the information given in time traces. Our article addresses this question.

2 Method and Results

Prima facie, the estimation of a mathematical model from time traces is a straightforward task of regression: Minimize the distance between the model and the data, and the resulting best-fitting model is characterized by a set of optimal parameter values that can be interpreted in the language of biology. Three complications render this task far from trivial. First, the regression requires a well-suited mathematical model. If one assumes a linear model with a known number of relevant variables, then the estimation of parameter values is an immediate matter of multivariate linear regression. However, very few phenomena in biology are linear, and as soon as nonlinearities begin to dominate, there is no longer guidance as to what model structure should be used for the regression. In metabolic studies, one might resort to traditional kinetic rate functions, such as the Michaelis-Menten rate law, but it has become evident in recent years that this type of function is not necessarily the best choice *in vivo* [4, 13]. Moreover, any change in this rate function, like taking inhibition into account, poses new questions of which type of inhibition or alteration would be most appropriate.

Second, the necessary nonlinearities greatly complicate any regression task, because there is seldom an analytical solution to the error-minimization problem [8]. Instead, one is forced to use a search algorithm, which often runs into convergence problems because of local minima, in which iterative searches may get stuck, and an error surface with unknown and often complex structure. Unless good initial guesses of most or all parameter values are at hand, one must expect difficulties with such searches if the biological system is of realistic size.

Third, very many relevant models in biology are based on systems of differential equations. Even if these are ordinary, they require costly numerical integration, which may dramatically slow down iterative searches, especially if the equations are stiff.

We propose to address these issues with a method that is based on three components. For the mathematical representation of biological systems we use *Biochemical Systems Theory* (*BST*; [10, 11]), which has proven valid and efficient for a large number of applications in a variety of biomedical research areas. While it is impossible to circumvent the issue of nonlinearities in searches for optimal parameter sets, we propose the preprocessing of data with an Artificial Neural Network (ANN), which is used to smooth the data and, combined with a stepwise regression, provides reasonable initial values and, ultimately, useful and relatively fast solutions. The costly evaluation of differential equations is circumvented with the replacement of derivatives with slopes, which are also estimated, algebraically, from the ANN-preprocessed data.

2.1 Biochemical Systems Theory (BST)

BST provides a general framework for modeling and analyzing nonlinear systems. It is based on the generic approximation of processes with products of power-law functions, a representation that is directly derived from multivariate linearization in logarithmic coordinates. This type of representation is compatible with the ubiquitous observation of allometry in biological systems and has been shown to be a valid description of biological processes in a variety of settings. Of importance here, BST models are easily scaled up to large sizes. Several-hundred journal articles, book chapters, full-size reviews and books [12, 17, 18, 19] have addressed various mathematical aspects of this formalism and its usefulness for theoretical and applied questions. It is therefore sufficient to limit the presentation of BST to a nutshell description.

System models in BST consist of sets of ordinary differential equations in which the change in each variable is always represented as sums and differences of multivariate products of power-law functions. The two most important variants within BST are the Generalized Mass Action form and the S-system form. In the former, each process is represented individually with a product of power-law functions and all these products are summed to capture the dynamics of a variable, while the latter always consists of only one difference between two power-law terms, one for all “incoming” fluxes and a second one for all “outgoing” fluxes. The generic S-system equation in this case reads:

$$dX_i/dt = \dot{X}_i = V_i^+ - V_i^- \approx \alpha_i X_1^{g_{i1}} X_2^{g_{i2}} \dots X_n^{g_{in}} - \beta_i X_1^{h_{i1}} X_2^{h_{i2}} \dots X_n^{h_{in}} \tag{1}$$

The structural simplicity of this type of equation permits a number of mathematical analyses, such as the algebraic determination and optimization of steady states, which in most other nonlinear models can only be obtained algorithmically with iterative methods that are costly and time-consuming [11, 17].

Of premier importance in the present context is that the parameters in BST models map directly onto interpretable biological features. For instance, the strength of inhibition exerted by X_1 on the production of X_2 is represented uniquely by the “kinetic order” parameter g_{21}. Since g_{21} represents an inhibition, it carries a negative value. If g_{21} is close to zero, the inhibition is weak, and with increasing magnitude, the strength of inhibition grows. The mapping between structural features and numerical quantities is a key advantage in the present context, because strong, weak, and no inhibition are special cases of the same mathematical function, which merely differ in the numerical value of g_{21}. The same is true for any other type of interaction between system components. As a crucial consequence, a machine-learning algorithm for structure identification can instead be trained to search merely for optimal sets of parameter values, because these correspond directly to the structure of the underlying S-system representation. By contrast, the analogous search in other nonlinear network models would require the learning algorithm to explore an unknown number of different mathematical forms, each time initiating a new search for suitable parameter values.

2.2 Data Preprocessing with Artificial Neural Networks (ANNs)

Parameter estimation tasks with S-systems or other nonlinear models are greatly facilitated if the data are noise-free. Of course, this situation is typically not realistic, but it can almost be achieved with a suitable preprocessing of the actual data. Specifically, it is beneficial to smooth the data in some relatively unbiased fashion and subsequently to use the smoothed data as substitutes for the actual, noisy data. Such smoothing is possible with different methods, including Kalman filters and splines, but we found it advantageous to use ANNs for this task as, if properly configured, they extract signal from noise directly from the experimental data and represent this signal algebraically [1, 7]. This is especially convenient for our purposes because it permits symbolic processing of the ANN-traced time series, such as the computation of slopes (see below). One must caution that any preprocessing of this type might bias the resulting estimates. This is unavoidable but also not very problematic, if these estimates are only used as initial guesses for a subsequent gradient search.

An ANN in our case consists of a multilayer perceptron with one input layer, one hidden layer, and one output layer ([24]; details are presented in [1], [7] and [21]). The value at each node h_j in the hidden layer is computed from the combination of m input values x_i with weights $w_{ij}^{<1>}$, according to a multivariate logistic function. Similarly, the values y_k of the output layer are determined as combinations of the l hidden node values h_j with weights $w_{jk}^{<2>}$, according to a second multivariate logistic function of the same type as before. While each individual function h_j and y_k is a simple sigmoid, the weighted summation and nesting of multiple of these sigmoids is able to produce functions of arbitrary flexibility, if the number of hidden nodes is sufficiently high [2]. The resulting output is therefore referred to as a "universal approximator" [5] or "universal function." Of importance is that even very large ANNs allow efficient analyses.

Smoothing the trace of metabolite X_i with an ANN amounts to determining a universal function that approximates the true solution $X_i(t)$ of the corresponding S-system differential equation, where the input data for the ANN are the measurement times. The output values are generated through "training" of the ANN, which consists of optimizing the numbers of hidden nodes and the weights $w_{ij}^{<1>}$ and $w_{jk}^{<2>}$ for both layers. The output that is practically more important for the user is a smoothed trace for each variable. These traces are internally represented by nested multivariate logistic functions, which allow the computation of slopes at any desired points. This computation is quite cumbersome, but can be executed straightforwardly with symbolic computer algebra software. Specific examples are shown in Section 2.4.

2.3 Decoupling of Differential Equations for the Purpose of Parameter Estimation

Given time traces and the symbolic form of an S-system as the alleged model, the identification of the structure of the underlying metabolic, genomic or proteomic system becomes a parameter estimation task or, as it is called in engineering, an "inverse" problem. In principle, this is a nonlinear regression task, but it is complicated by the fact that the model is described by a system of differential equations.

For the estimation purposes that are of interest here, it is possible to decouple the set of differential equations by replacing all derivatives on the left-hand sides of the equations with slopes, which are to be estimated directly from the observed data. We use for this estimation the ANN method described in Section 2.2. Thus, given trace points $X_k(t_j)$ and slopes $S_i(t_j)$ at N time points t_j, the estimation is performed with a set of $n \times N$ algebraic equations, where n is the number of differential equations in the original system. Each equation of this set has the form

$$S_i(t_j) \approx \alpha_i X_1^{g_{i1}}(t_j) X_2^{g_{i2}}(t_j) \dots X_n^{g_{in}}(t_j) - \beta_i X_1^{h_{i1}}(t_j) X_2^{h_{i2}}(t_j) \dots X_n^{h_{in}}(t_j) \quad (2)$$

These algebraic equations are not solutions of the differential equations (1) in the traditional sense, and one could only use them in an approximate, iterative sense to compute dynamic time trends, as Euler's method does it for numerically integrating differential equations. However, for purposes of parameter estimation, the algebraic equations are perfect analogues within the accuracy of trace and slope estimation. They generate constraints within the parameter space that ideally are so tight that only one (the true) solution is able to satisfy them. One may intuit the validity of this conversion of coupled differential equations into decoupled algebraic equations by realizing that the values of the system variables (X's) at different times become the data points in the regression problem and that the parameters (a's, b's, g's and h's) become the regression "variables."

By converting the system of differential equations into a larger set of algebraic regression equations, the inverse problem is greatly simplified, because it does no longer require costly numerical integration, which becomes particularly problematic if the differential equations are stiff [21]. Furthermore, the decoupling allows analysis of one equation at a time. This fact is not simply convenient computationally, for instance, for direct parallelization, but has the advantage that estimation errors in one equation do not influence parameter estimates in the other equations.

2.4 Example: Ethanol Fermentation in Yeast

As an illustration of the method, we re-analyze experimental data on the batch fermentation of glucose in the baker's yeast *Saccharomyces cerevisiae* [22]. Four variables had been measured over an 11-hour time horizon, namely the concentrations of glucose and ethanol, and the biomass of viable and nonviable cells. The latter two were also recorded as total biomass in the system. At the beginning of the experiment, glucose was added as a one-time bolus and subsequently used up by the yeast cells for the production of ethanol. Thus, the glucose continuously decreased in concentration, while the ethanol concentration monotonically increased. The number of viable cells was initially found to increase slightly, but as glucose was being used up, the population growth slowed down and soon the number of viable cells began to decrease until essentially all cells were no longer viable. The original data are shown in Fig. 1, superimposed with ANN-computed traces from our webtool *Webmetabol* (`https://bioinformatics.musc.edu/webmetabol/`).

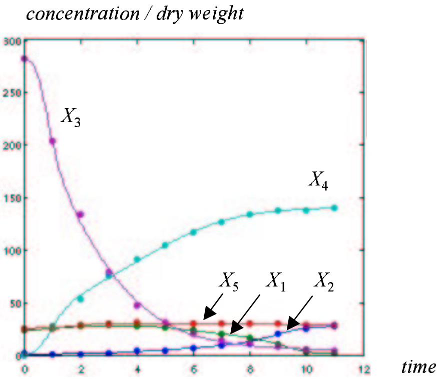

Figure 1: *Webmetabol* snapshot of fermentation data [22], superimposed with ANN-derived traces. Labels and arrows were added to facilitate trace identification. See Text for variable names.

While the interactions between the system variables are relatively clear, there are some questions about ethanol catabolism and the possible lysis of nonviable cells. Thus, some information about this physiological pathway is available, but this information is incomplete; the tentative diagram of the system is given in Fig. 2.

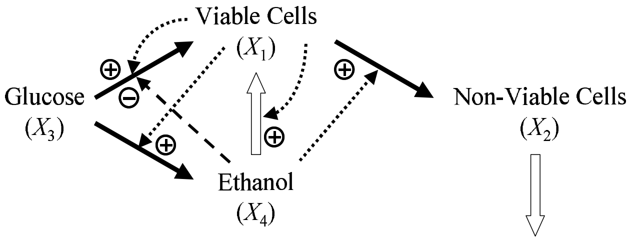

Figure 2: Diagram used to describe biotechnological data on ethanol fermentation. Solid arrows show flow of material, empty arrows show putative flow (cell lysis and ethanol catabolism). Dotted arrows indicate activating effects and dashed arrow represents putative inhibition.

The construction of the S-system model is straightforward. Four equations are constructed for the main variables, namely, viable cells (X_1), nonviable cells (X_2), glucose (X_3), and ethanol (X_4). Total biomass (X_5) does not require its own equation, because it is simply the sum of the first two variables. Without any further information, each α- and β-term would contain a rate constant and all four variables with their respective kinetic orders, thus requiring the estimation of 40 parameters. The naïve, straightforward way of analyzing the data with this model would thus consists of smoothing the traces with the ANN, computing slopes from the traces, and regressing the model. The idealistic result would be the true set of parameter values. This strategy is not preferable, because it is to be expected from prior experience (*e.g.*, [15]) that many combinations of parameter values could fit this single set of data equally well. In cases where several datasets are available, the space of possible solutions is drastically constrained [6], but in the present case, additional data are not at hand.

Fortunately, BST allows us to convert additional information about the pathway into parameter values or into bounds on parameter values. Thus, the most general symbolic S-system can be greatly reduced by constraints derived immediately from the biology of the system. As a rather obvious example, ethanol is not converted into glucose under the given conditions. In fact, there is no glucose production at all in this bolus experiment, so that the corresponding equation (of X_3) consists only of a consumption term. Even this consumption term can be simplified, because neither ethanol nor the nonviable cells are likely to affect this term directly. Thus, the third equation is reduced in complexity from

$$\dot{X}_3 = \alpha_3 X_1^{g_{31}} X_2^{g_{32}} X_3^{g_{33}} X_4^{g_{34}} - \beta_3 X_1^{h_{31}} X_2^{h_{32}} X_3^{h_{33}} X_4^{h_{34}} \tag{3}$$

to

$$\dot{X}_3 = -\beta_3 X_1^{h_{31}} X_3^{h_{33}} labeleq4 \tag{4}$$

This remarkable simplification is facilitated by two circumstances. First and rather obvious, the additional biological information allows us to constrain the possibilities afforded by the generic Eq. (3). Second, the particular structure of S-systems is based on a one-to-one relationship between parameter values and processes. For instance, the kinetic order h_{34} would refer to the direct effect of ethanol on glucose consumption. The absence of this interaction mandates directly and unambiguously that h_{34} be zero, with the consequence that X_4 effectively drops out of the equation. This advantage needs to be seen in comparison, for instance, with a polynomial representation, where a change in one component of the system often affects many or all coefficients of the polynomial.

In contrast to the simplification of the third equation, the first equation, describing the population dynamics of viable cells, needs to be kept fairly general for the estimation. For instance, it is clear that the cells contribute to their own population growth by virtue of cell division. Furthermore, glucose is the growth substrate and therefore should contribute to the increase in cell numbers. Finally, one could surmise that ethanol could inhibit growth, contribute to cell death, or, on the contrary, be used as an additional substrate. Because these effects are not *a priori* obvious, the ethanol variable X_4 is kept in both, the production and the degradation terms for the viable cells. Still, simplifications are possible. For instance, nonviable cells do not contribute to the dynamics of the viable cell population; their pool is simply a passive recipient. Also, glucose is not likely to contribute to the decrease in the population. Again, these singular pieces of information are directly convertible into setting parameter values equal to zero.

The nonviable cells arise from viable cells, and this killing could possibly be affected by ethanol, but probably not by glucose. Thus, the production term contains X_1 and X_4. Furthermore, cell lysis should be retained as a possible mechanism, because the total biomass is decreasing toward the end of the experiment. Thus, X_2 needs to be included in the degradation term of X_2. If cell lysis is not of importance, the regression should identify the corresponding rate constant b_2 as zero or as statistically insignificant.

Finally, the production term in the ethanol equation is affected by glucose availability and the number of cells. There is no obvious route of degradation of ethanol. However, to test the hypothesis that ethanol might be catabolized by the cells for growth, a degradation term should be included containing ethanol itself and the number of viable cells. Thus, the reduced equations in symbolic form read

$$\begin{aligned}
\dot{X}_1 &= \alpha_1 X_1^{g_{11}} X_3^{g_{13}} X_4^{g_{14}*} - \beta_1 X_1^{h_{11}} X_4^{h_{14}} \\
\dot{X}_2 &= \beta_1 X_1^{h_{11}} X_4^{h_{14}} - \beta_2^* X_2^{h_{22}} \\
\dot{X}_3 &= -\beta_3 X_1^{h_{31}} X_3^{h_{33}} \\
\dot{X}_4 &= \alpha_4 X_1^{g_{41}} X_3^{g_{43}} 4 - \beta_4^* X_1^{h_{41}} X_4^{h_{44}}
\end{aligned} \tag{5}$$

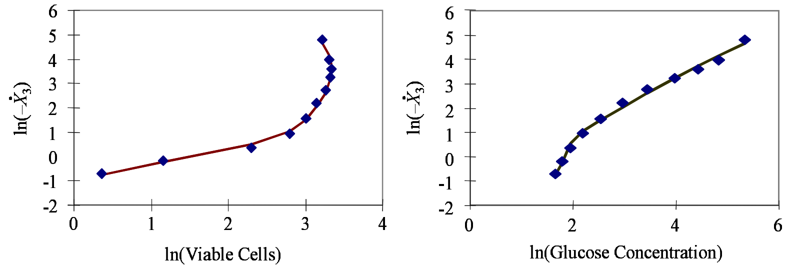

Figure 3: Projection plots of the (negative) slope in the glucose equation against variables directly affecting the dynamics, in logarithmic coordinates. Symbols show the smoothed data points, which in the case essentially coincide with original data (*cf.* Fig. 1).The lines were computed through linear regression from ANN-determined trace points and slopes.

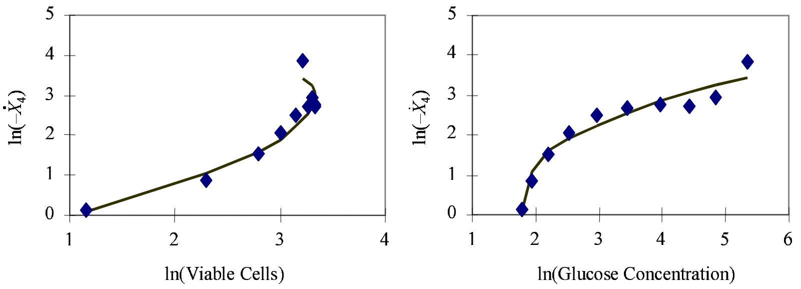

Figure 4: Projection plots of the slope in the ethanol equation against variables directly affecting the dynamics, in logarithmic coordinates.

where the asterisks denote parameters that, according to *a priori* information, might be zero.

The first step of the analysis consists of training the ANN on the data, for which we developed specific code and configured it as a web-based application [24]. The training is done initially through fully automated selection of hidden nodes, epochs, and cross-validation, and may be followed up with additional fitting, using the same web-based application, where the user suggests different numbers of hidden nodes and epochs. The result is shown in Fig. 1, which, except for labeling, is taken unaltered from the web-interface. The fit is generally good, but one may note that the shapes of the traces at the very beginning and end of the datasets are not always consistent with our expectation. A sigmoidal start appears to be a general tendency with ANN smoothing, and it is useful to run regressions with and without these extreme points.

The next step is the automatic computation of slopes from the ANN traces, which is also performed with the same web-based interface. The slopes may be computed just at the observed data points or anywhere along the trace. The results are available in numerical form for further internal evaluation or may be exported to other programs like Excel for further analysis.

As an example, given the slope estimates of the ANN, the third equation in (5) becomes linear upon logarithmic transformation and can directly be regressed in a convenient spreadsheet application. As three alternatives, the regression may be executed with all data points, with all except for the first and last data points, or with other trace points and slopes that are directly obtained from the ANN-smoothed traces. There is no restriction to using one or another alternative for different traces. The result for the third equation (using all data without extremes) is shown in Fig. 3, where the (logarithm of the) slope of the glucose dynamics is plotted against (the logarithm of) glucose and (the logarithm of) the numbers of cells, respectively; since both sides are entirely in the negative domain, they were first multiplied by -1. The plots are shown as projections in logarithmic coordinates, where the goodness of fit is characterized by $R^2 = 0.996$. This quality indicates that glucose consumption is modeled well. The corresponding S-system parameters are $b_3 = 0.0659$, $h_{31} = 0.49$, and $h_{33} = 1.1$.

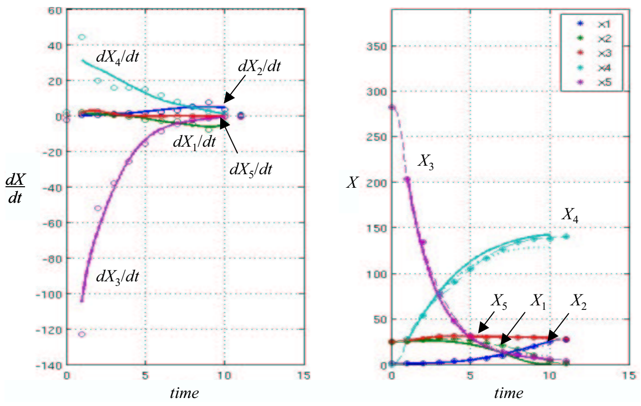

Figure 5: Graphical *Webmetabol* output. Left panel: Data fit based on slopes. Right panel: Time courses.

The dynamics of ethanol was hypothesized either to consist just of the production term or of production and catabolism. We first test the simpler case where catabolism does not exist or is negligible. Again, the regression is linear in logarithmic coordinates. The results are shown in Fig. 4. In this case, $R^2 = 0.945$, which again is a good fit, even though one detects some possibly systematic deviation between the prediction and the observed glucose values. The corresponding S-system parameters are $a_4 = 0.185$, $g_{41} = 0.80$, and $g_{43} = 0.48$. Regression including both production and catabolism cannot be done with linear methods and requires an iterative search, which is facilitated by the webtool [24]. The result is a slight improvement in fit, which however does not seem to be significant, although no formal statistical test was performed (cf. [25] for such a test).

Webmetabol [24] can be used to fit all traces against a full S-system model, but it also allows fixing those parameters that are known. In particular, if interactions are known not to exist, the corresponding kinetic orders and/or rate constants are forced to be 0. Using the reduced equations in (5), *Webmetabol* produces the fit in Fig. 5; again, labels were added. The left panel shows the regression using the ANN-derived traces and slopes, and the right panel shows the corresponding dynamics. One notes that the time courses of the viable cells and the ethanol production are slightly shifted. This is due to the fact that *Webmetabol* uses a data point as initial value, without additional optimization. Slight changes in these initial values shift these curves into the right positions. While the data fit is obviously of interest, an interpretation of the estimated kinetic orders is equally important.

The production dynamics of viable cells according to the results depends on the present number of cells, but not on glucose and ethanol. In the earlier analysis [22], which produced a data fit of similar quality, cell population growth depended on glucose also. The analysis thus yields two alternative solutions between which one can only select with further data. While the dependence on glucose in the earlier analysis had been taken for granted, the results here would suggest that during population growth–*i.e.*, during the first four hours–glucose is plentiful and not limiting. Subsequently, the population dynamics is dominated by cell death.

According to this and the earlier analysis citebib22, cell death is strongly affected by ethanol. In both cases the kinetic order is surprisingly high (2.65 [22] versus 3.07 here). Also as previously, accounting for cell lysis improves the data fit, especially toward the end of the experiment.

The glucose equation was fitted per linear regression, as discussed above. *Webmetabol* obtains a structurally equivalent and numerically similar equation. The ethanol dynamics is structurally equivalent to the earlier results, indicating that the number of viable cells and the glucose concentration are important. Also as in [22], the fit is slightly improved if ethanol degradation is accounted for. In both cases, the number of viable cells and ethanol are the two components of the degradation term, and in both cases, the flux through this term is very small, suggesting secondary importance of ethanol utilization.

3 Discussion

Time traces, whether describing gene expression, gene regulation, protein prevalence, metabolite concentrations, or a mix of biochemical and physiological data as in the example presented here, contain a wealth of information on the underlying organization and regulation of the biological system. This information is implicit, and the challenge is to retrieve it efficiently and without too much bias. In principle, this retrieval poses a regular inverse problem, but the fact that models for biological phenomena are almost always nonlinear and based on differential equations makes the seemingly straightforward task difficult [7]. The methods proposed here do not solve all issues associated with the retrieval of information from time traces. However, they greatly simplify this process through the choice of a convenient model structure, namely BST and S-systems, the preprocessing of data with a neural network, and a "trick" that circumvents the integration of the differential equations and allows the investigation of traces one at a time. Technical details of this combination of steps are described elsewhere [21], and the emphasis here was on the demonstration of how additional pieces of information can be merged with the information obtainable from time traces. Together, this information effectively constrains the space of possible parameter values that are consistent with observations. Once estimates or ranges of these parameters are established, they can be interpreted within the realm of the application area.

The methods were illustrated with a specific example for which the true parameter values are not known. Earlier estimations yielded numerical values slightly different from those obtained here, which is not too surprising, because only one dataset was available for analysis. Importantly, however, both analyses led to consistent interpretations. Even more important is that the example may serve as a proof of concept for a novel strategy for solving inverse problems. This strategy combines a suitable modeling framework with methods for data smoothing and slope estimation, which in turn circumvents the need to solve differential equations. Each of these methods can be scaled up to realistically sized networks and therefore ameliorates the "curse of dimensionality," which accompanies most inverse problems. Of course, noise in each step confounds noise in previous steps, so that it remains to be seen how scaleable the method is. It is also clear that large systems will allow for wider latitude in possible parameter combinations that fit experimental data equally well. It seems that this problem can only be overcome by the availability of multiple datasets, additional information on the connectivity of the network, and preprocessing that allows starting the unavoidable regression closer to the true solution. Without such helpful input, the proposed algorithm typically leads to disparate sets of parameter values, each with their own interpretation. Additional data sets will reduce this range of possibilities, ideally to a single organizational and regulatory structure. But even if such data are scarce, it was shown here that biological insight can be utilized with similar efficiency. While the proposed prototype method shows great potential, much research will be needed to optimize its details.

Acknowledgments

This work was supported in part by a Quantitative Systems Biotechnology grant (BES-0120288; E.O. Voit, PI) from the National Science Foundation, a Cancer Center grant from the Department of Energy (C.E. Reed, PI), a National Heart, Lung and Blood Institute Proteomics Initiative through contract N01-HV-28181; D. Knapp, PI), and an Interdisciplinary USC/MUSC grant (E.P. Gatzke, PI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring institutions.

References

[1] Almeida, J.S., Predictive non-linear modeling of complex data by artificial neural networks, *Curr. Opin. Biotechn.*, 13(1):72–76, 2002.

[2] Funahashi, K.-I., On the approximate realization of continuous mappings by neural networks, *Neural Networks*, 2:183–192 , 1989.

[3] Goodacre, R. and Harrigan, G.G. (Eds), *Metabolite Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, Kluwer Academic Publishing, Dordrecht, The Netherlands, 2003.

[4] Hill, C.M., R.D. Waight, and Bardsley, W.G., Does any enzyme follow the Michaelis-Menten equation?, *Molec. Cell. Biochem.*, 15:173–178, 1977.

[5] Hornik, K., Stinchcombe, M., and White, H., Multilayer feedforward networks are universal approximators, *Neural Networks*, 2:359–366, 1989.

[6] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., and Tomita, M., Dynamic modeling of genetic networks using genetic algorithm and S-system, *Bioinformatics*, 19(5):643–650, 2003.

[7] Mendes, P. and Kell, D.B., On the analysis of the inverse problem of metabolic pathways using artificial neural networks, *BioSystems*, 38:15–28, 1996.

[8] Mendes, P. and Kell, D.B., Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation, *Bioinformatics*, 14:869–883, 1998.

[9] Neves, A.R., Ventura, R., Mansour, N., Shearman, C., Gasson, M.J., Maycock, C., Ramos, A., and Santos, H., Is the glycolytic flux in *Lactococcus lactis* primarily controlled by the redox charge?, *J. Biol. Chem.*, 277(31):28088–28098, 2002.

[10] Savageau, M.A., Biochemical Systems Analysis, I. Some mathematical properties of the rate law for the component enzymatic reactions, *J. Theor. Biol.*, 25:365–369, 1969.

[11] Savageau, M.A., Biochemical Systems Analysis, II. The steady-state solutions for an n-pool system using a power-law approximation, *J. Theor. Biol.*, 25:370–379, 1969.

[12] Savageau, M.A., *Biochemical Systems Analysis. A Study of Function and Design in Molecular Biology*, Addison-Wesley, Reading, Massachusetts, 1976.

[13] Savageau, M.A., Enzyme kinetics *in vitro* and *in vivo*: Michaelis-Menten revisited, In Bittar, E.E., (Ed.), *Principles of Medical Biology*, 4:93–146, JAI Press Inc., Greenwich, Conneticut, 1995.

[14] Sorribas, A. and Cascante, M., Structure identifiability in metabolic pathways: parameter estimation in models based on the power-law formalism, *Biochem. J.*, 298:303–311, 1994.

[15] Sorribas, A., March, J., and Voit, E.O., Estimating age-related trends in cross-sectional studies using S-distributions, *Stat. in Med.*, 10(5):697–713, 2000.

[16] Theobald, U., Mailinger, W., Baltes, M, Rizzi, M., and Reuss, M., *In vivo* analysis of metabolic dynamics in *Saccharomyces cerevisiae*: I. Experimental observations, *Biotechn. Bioeng.*, 55(2):305–316, 1997.

[17] Torres, N.V. and Voit, E.O., *Pathway Analysis and Optimization in Metabolic Engineering*, Cambridge University Press, Cambridge, U.K., 2002.

[18] Voit, E.O. (Ed.), *Canonical Nonlinear Modeling. S-System Approach to Understanding Complexity,* xi+365 pp., Van Nostrand Reinhold, New York, 1991.

[19] Voit, E.O., *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists*, xii + 530 pp., Cambridge University Press, Cambridge, U.K., 2000.

[20] Voit, E.O., Models-of-data and models-of processes in the post-genomic era. Special Issue in honor of John A. Jacquez, *Math. Biosc.*, 180:263–274, 2002.

[21] Voit, E.O. and Almeida, J.S., Decoupling dynamical systems for pathway identification from metabolic profiles, submitted.

[22] Voit, E.O. and Savageau, M.A., Power-law approach to modeling biological systems; II. Application to ethanol production, *J. Ferment. Technol.*, 60(3):229–232, 1982.

[23] Voit, E.O. and Savageau, M.A., Power-law approach to modeling biological systems; III. Methods of analysis, *J. Ferment. Technol.*, 60(3):233–241, 1982.

[24] `https://bioinformatics.musc.edu/webmetabol/`

[25] `http://rana.lbl.gov/EisenData.htm`